Vol.2019-SLP-126 No.9
2019/2/28

正弦関数摂動 von Mises 分布 DNN のモード近似を用いた位相復元

高道 慎之介[1,a)]　猿渡 洋[1,b)]

概要：統計的音声音響信号処理では，短時間フーリエ変換の振幅スペクトログラムに対する処理がしばしば行われる. 最終的な音声波形を生成するためには，振幅スペクトログラムから位相情報を復元する必要がある．これに対し我々は，von Mises 分布を条件付き確率分布として有する von Mises 分布 deep neural network (DNN) に基づく位相・群遅延モデリングを提案し，DNN 学習と位相推定で統一された目的関数を導入した．更に我々は，正弦関数摂動 von Mises 分布 DNN により，群遅延ヒストグラムのような条件付き非対称周期分布のモデル化を可能にした．しかしながら，正弦関数摂動 von Mises 分布のモード（最頻値）は解析的に求まらないため，位相と群遅延の両方を考慮した目的関数を設計しづらい．そこで本稿では，当該分布のモードを微分可能かつ解析的な形で近似することで，統一された目的関数による学習・推定を可能にする．実験的評価により提案法の有効性を示す．

Shinnosuke Takamichi[1,a)]　Hiroshi Saruwatari[1,b)]

1. はじめに

音源分離や音声強調などの音響信号処理ではしばしば，短時間フーリエ変換 (short-term Fourier transform: STFT) による振幅スペクトログラムに対する処理が行われる．また，近年の統計的音声合成 [1] は，ボコーダパラメータを生成する枠組みから振幅スペクトルを直接的に生成する枠組み [2], [3], [4] に移行しつつある．これらの技術により最終的な音声を生成する場合，与えられた振幅スペクトログラムに対応する位相スペクトログラムが必要だが，その位相スペクトログラムは得られない場合が多い．

これに対し我々は，deep neural network (DNN) を用いた生成モデル（深層生成モデル）に基づく位相推定法を提案した．von Mises 分布 DNN [5] は，条件付き確率分布として von Mises 分布 [6] を有する深層生成モデルであり，位相のような周期変数のモデル化に適している．我々はこれを用いて，振幅スペクトログラムからの位相推定法を提案した [5]．この枠組みでは，位相と群遅延の両方を考慮した統一的な目的関数により，DNN 学習と位相推定が可能である．一方で，von Mises 分布のような対称分布では，非対称に分布する群遅延のヒストグラムを効率的にモデル化できない．これに対し我々は，正弦関数摂動 von Mises 分布 DNN を提案し，群遅延ヒストグラムをより効率的にモデル化できることを明らかにした [7]．しかしながら，正弦関数摂動 von Mises 分布のモード（最瀕値）は解析的に求まらないため，位相と群遅延の両方を考慮した統一的な目的関数の設計が困難である．

そこで本稿では，正弦関数摂動 von Mises 分布のモード近似法と，正弦関数摂動 von Mises 分布 DNN を用いた位相・群遅延モデリングを提案する．これまで反復法により推定していたモードを解析的かつ微分可能な形で近似することで，位相と群遅延のマルチタスクの形式で DNN 学習・位相推定のための統一された目的関数を設計できる．実験的評価から，提案法は，従来の von Mises 分布 DNN に基づく手法よりも，完全再構成に近い位相を生成できることを示す．

2. von Mises 分布 DNN を用いた位相・群遅延モデリング [5]

振幅スペクトログラムからの位相復元のための，von Mises 分布 DNN を用いた位相・群遅延モデリングを概説する．ここで，$\boldsymbol{x} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T]$ と $\boldsymbol{y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t, \cdots, \boldsymbol{y}_T]$ をそれぞれ，振幅・位相スペ

[1] 東京大学 大学院情報理工学系システム情報学専攻, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[a)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[b)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

ⓒ 2019 Information Processing Society of Japan

クトログラムとする．$\boldsymbol{x}_t = [x_{t,0}, \cdots, x_{t,f}, \cdots, x_{t,F}]^\top$ と $\boldsymbol{y}_t = [y_{t,0}, \cdots, y_{t,f}, \cdots, y_{t,F}]^\top$ はそれぞれ，時刻 t における振幅及び位相である．f は周波数ビンのインデックスであり，F はナイキスト周波数に対応する．$x_{t,f}$ と $y_{t,f}$ は実数値であり，$y_{t,f}$ は，2π の周期をもつ周期変数である．DNN 学習と位相推定は，位相・群遅延が時間周波数依存の von Mises 分布に従うと仮定したもとでの最尤推定に基づき行われる．

2.1 von Mises 分布

von Mises 分布 $P^{(\mathrm{vm})}(y; \mu, \kappa)$ を，次式で定義する．

$$P^{(\mathrm{vm})}(y; \mu, \kappa) = \frac{\exp(\kappa \cos(y - \mu))}{2\pi I_0(\kappa)} \tag{1}$$

ここで，$I_0(\cdot)$ は 0 次の第 1 種変形 Bessel 関数である．μ は平均，κ は，ガウス分布の精度（分散の逆数）に対応する集中度パラメータであり，非負値をとる．本来，von Mises 分布は多次元変数をとらないが，本稿では表記の簡単化のため，スカラ変数に対する von Mises 分布の積を

$$P^{(\mathrm{vm})}(\boldsymbol{y}_t; \boldsymbol{\mu}_t, \boldsymbol{\kappa}_t) = \prod_{f=0}^{F} P^{(\mathrm{vm})}(y_{t,f}; \mu_{t,f}, \kappa_{t,f}) \tag{2}$$

と表記する．ここで，$\boldsymbol{\mu}_t = [\mu_{t,0}, \cdots, \mu_{t,f}, \cdots, \mu_{t,F}]^\top$ と $\boldsymbol{\kappa}_t = [\kappa_{t,0}, \cdots, \kappa_{t,f}, \cdots, \kappa_{t,F}]^\top$ はそれぞれ，$\boldsymbol{y}_t$ に対応する平均ベクトルと集中度パラメータベクトルである．

2.2 群遅延

群遅延は，位相の周波数微分の負値として定義される．本稿では，群遅延を一次差分で近似する．

$$\Delta y_{t,f} = -(y_{t,f+1} - y_{t,f}) \tag{3}$$

位相 $y_{t,f}$ は周期変数であるが，群遅延 $\Delta y_{t,f}$ もまた周期変数である．以降では，時刻 t における群遅延を $\Delta \boldsymbol{y}_t = [\Delta y_{t,0}, \cdots, \Delta y_{t,f}, \cdots, \Delta y_{t,F}]^\top$ とする．この位相−群遅延変換は，次のように行列表現も可能である．

$$\Delta \boldsymbol{y}_t = \boldsymbol{W} \boldsymbol{y}_t \tag{4}$$

$$\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_f, \cdots, \boldsymbol{w}_{F+1}]^\top \tag{5}$$

$$\boldsymbol{w}_f = \Big[\underset{1\mathrm{st}}{0}, \cdots, 0, \underset{f\mathrm{th}}{1}, -1, 0, \cdots, \underset{(F+1)\mathrm{th}}{0}\Big]^\top \tag{6}$$

この行列 $\boldsymbol{W}$ は逆行列をもつ (具体的には，非零の要素が 1 の下三角行列) ため，$\boldsymbol{y}_t = \boldsymbol{W}^{-1} \Delta \boldsymbol{y}_t$ による逆変換も可能である．

2.3 最大化する目的関数

DNN 学習と位相推定で統一された次式の目的関数を用いる．なお，添字の ph+gd は，位相 (phase) と群遅延 (group delay) の両者を考慮していることを表す．

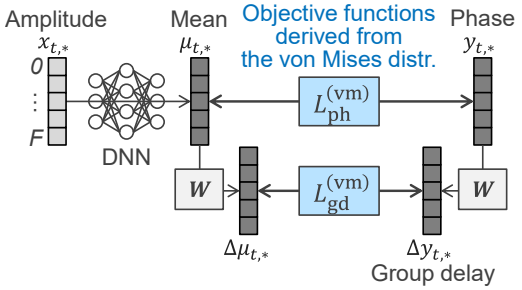

図 1 von Mises 分布 DNN を用いた位相・群遅延モデリングのアーキテクチャ．簡単化のため，この図ではフレーム毎に振幅から位相を推定しているが，実際の実験的評価では，マルチフレームの振幅から位相を推定している．

$$L_{\mathrm{ph+gd}}^{(\mathrm{vm})} = L_{\mathrm{ph}}^{(\mathrm{vm})} + \alpha_{\mathrm{gd}} L_{\mathrm{gd}}^{(\mathrm{vm})} \tag{7}$$

$$L_{\mathrm{ph}}^{(\mathrm{vm})} = -\log P^{(\mathrm{vm})}(\boldsymbol{y}_t; \boldsymbol{\mu}_t, \kappa \mathbf{1}) \tag{8}$$

$$\begin{aligned} L_{\mathrm{gd}}^{(\mathrm{vm})} &= -\log P^{(\mathrm{vm})}(\Delta \boldsymbol{y}_t; \Delta \boldsymbol{\mu}_t, \kappa \mathbf{1}) \\ &= -\log P^{(\mathrm{vm})}(\boldsymbol{W} \boldsymbol{y}_t; \boldsymbol{W} \boldsymbol{\mu}_t, \kappa \mathbf{1}) \end{aligned} \tag{9}$$

ここで，$\mathbf{1}$ は全ての要素が 1 のベクトル，κ は時間周波数非依存の定数値，α_{gd} は群遅延重みである．$\Delta \boldsymbol{\mu}_t = [\Delta \mu_{t,0}, \cdots, \Delta \mu_{t,f}, \cdots, \Delta \mu_{t,F}]^\top$ は時刻 t における群遅延の平均ベクトルである．

2.4 学習

DNN のモデルパラメータを $\boldsymbol{\theta}_{\mathrm{G}}$ とする．この DNN は，図 1 に示すように，時刻 t において，入力 $\boldsymbol{x}$ から位相の平均ベクトル $\boldsymbol{\mu}_t$ を出力する．その後，$\boldsymbol{W}$ による線形変換から群遅延の平均ベクトル $\Delta \boldsymbol{\mu}_t = \boldsymbol{W} \boldsymbol{\mu}_t$ を得る．$\boldsymbol{\theta}_{\mathrm{G}}$ は，次式に示すように，式 (7) の負の対数値を最小化するように推定される．

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{G}} = \underset{\boldsymbol{\theta}_G}{\mathrm{argmin}} \frac{1}{T} \sum_{t=1}^{T} L_{\mathrm{ph+gd}}^{(\mathrm{vm})} \tag{10}$$

ここで，式 (7) の負の対数値は次式で与えられる．

$$\begin{aligned} L_{\mathrm{ph+gd}}^{(\mathrm{vm})} = &- \sum_{f=0}^{F} \cos(y_{t,f} - \mu_{t,f}) \\ &- \alpha_{\mathrm{gd}} \sum_{f=0}^{F} \cos(\Delta y_{t,f} - \Delta \mu_{t,f}) \end{aligned} \tag{11}$$

2.5 推定

学習済み DNN を用いて位相を推定する．時刻 t における推定位相を $\hat{\boldsymbol{y}}_t$ とすると，$\hat{\boldsymbol{y}}_t$ は次式で得られる．

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}_t}{\mathrm{argmin}}\, L_{\mathrm{ph+gd}}^{(\mathrm{vm})} = \boldsymbol{\mu}_t \tag{12}$$

すなわち，学習済み DNN の出力 $\boldsymbol{\mu}_t$ が，直接的に推定位相となる．以上より，式 (7) は，学習・推定時に矛盾のない統一された目的関数であることが分かる．

ⓒ 2019 Information Processing Society of Japan

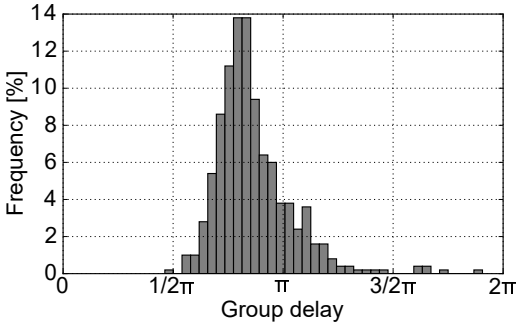

図 2　周波数 1 kHz における群遅延ヒストグラムの例．振幅スペクトルにより条件づけられたヒストグラムを描画するため，ここでは，k 近傍の振幅スペクトルに対応する群遅延を用いた．

2.6　問題点

群遅延は，周波数スペクトルの極・零点によって，線形位相成分から正または負の方向に変化する．また，自己回帰モデルを仮定した場合，群遅延は負よりも正の方向に変化するため，群遅延ヒストグラムは非対称となる．図 2 に，式 (3) で計算した群遅延の条件付きヒストグラムの例を示す．前述したように，ヒストグラムは非対称であることが確認できる．von Mises 分布は対称分布であるため，von Mises 分布 DNN は，このような非対称ヒストグラムのモデル化に適さない．

3.　正弦関数摂動 von Mises 分布 DNN を用いた群遅延モデリング [7]

3.1　正弦関数摂動 von Mises 分布

正弦関数摂動 von Mises 分布（図 3） [8] は，von Mises 分布に摂動項をかけた次式で定義される．

$$P^{(\mathrm{ssvm})}\left(y;\mu,\kappa,\lambda\right) = P^{(\mathrm{vm})}\left(y;\mu,\kappa\right) P^{(\mathrm{ss})}\left(y;\mu,\lambda\right) \quad (13)$$

$$P^{(\mathrm{ss})}\left(y;\mu,\lambda\right) = 1 + \lambda \sin\left(y-\mu\right) \quad (14)$$

λ は，摂動パラメータであり，$-1 \leq \lambda \leq 1$ の値をとる．本稿では表記の簡易化のため，摂動項を確率分布のように表記 ($P^{(\mathrm{ss})}\left(\cdot\right)$) するが，この項を全区間で積分すると 2π になることに注意する．また，当該分布は多変量変数をとらないが，本稿では表記の簡単化のため，スカラ変数に対する分布の積を

$$P^{(\mathrm{ssvm})}\left(\boldsymbol{y}_t;\boldsymbol{\mu}_t,\boldsymbol{\kappa}_t,\boldsymbol{\lambda}_t\right) = P^{(\mathrm{vm})}\left(\boldsymbol{y}_t;\boldsymbol{\mu}_t,\boldsymbol{\kappa}_t\right) P^{(\mathrm{ss})}\left(\boldsymbol{y}_t;\boldsymbol{\mu}_t,\boldsymbol{\lambda}_t\right) \quad (15)$$

$$P^{(\mathrm{ss})}\left(\boldsymbol{y}_t;\boldsymbol{\mu}_t,\boldsymbol{\lambda}_t\right) = \prod_{f=0}^{F} P^{(\mathrm{ss})}\left(y_{t,f};\mu_{t,f},\lambda_{t,f}\right) \quad (16)$$

と表記する．ここで，$\boldsymbol{\lambda}_t = \left[\lambda_{t,0},\cdots,\lambda_{t,f},\cdots,\lambda_{t,F}\right]^\top$ は，摂動パラメータベクトルである．

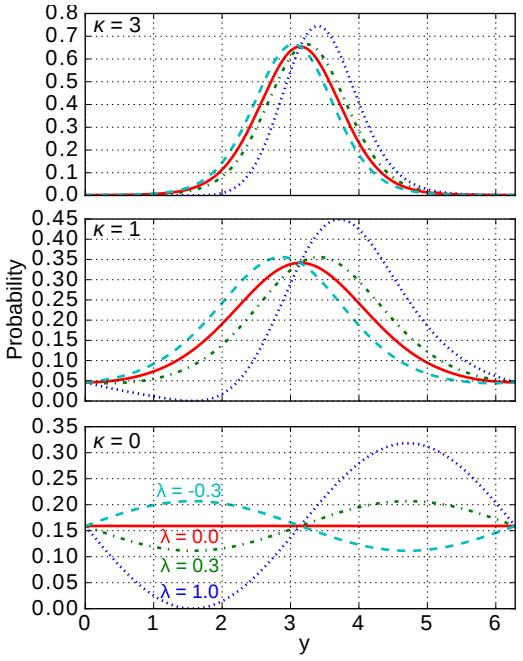

図 3　正弦関数摂動 von Mises 分布の例．$\mu = \pi$ としている．$\lambda = 0$ の場合，この確率分布は von Mises 分布と等価となる．

3.2　最大化する目的関数

群遅延が時間周波数依存の当該分布に従うと仮定し，次式の目的関数を用いる．

$$L_{\mathrm{gd}}^{(\mathrm{ssvm})} = -\log P^{(\mathrm{ssvm})}\left(\Delta\boldsymbol{y}_t;\boldsymbol{\mu}_t,\boldsymbol{\kappa}_t,\boldsymbol{\lambda}_t\right) \quad (17)$$

3.3　学習

モデルパラメータ $\boldsymbol{\theta}_{\mathrm{G}}$ をもつ DNN は，時刻 t において，入力 $\boldsymbol{x}$ から $\boldsymbol{\mu}_t,\boldsymbol{\kappa}_t,\boldsymbol{\lambda}_t$ を出力する．$\boldsymbol{\theta}_{\mathrm{G}}$ は，次式に示すように，式 (17) を最小化するように推定される．

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{G}} = \underset{\boldsymbol{\theta}_G}{\mathrm{argmin}} \frac{1}{T}\sum_{t=1}^{T} L_{\mathrm{gd}}^{(\mathrm{ssvm})} \quad (18)$$

$$L_{\mathrm{gd}}^{(\mathrm{ssvm})} = \sum_{f=0}^{F}\{\log\frac{I_0\left(\kappa_{t,f}\right)}{1+\lambda_{t,f}\sin\left(\Delta y_{t,f}-\mu_{t,f}\right)} - \kappa_{t,f}\cos\left(\Delta y_{t,f}-\mu_{t,f}\right)\} \quad (19)$$

3.4　推定

学習済み DNN を用いて群遅延を推定する．時刻 t における推定群遅延を $\Delta\hat{\boldsymbol{y}}_t$ とすると，$\Delta\hat{\boldsymbol{y}}_t$ は次式で得られる．

$$\Delta\hat{\boldsymbol{y}}_t = \underset{\Delta\boldsymbol{y}_t}{\mathrm{argmin}}\, L_{\mathrm{gd}}^{(\mathrm{ssvm})} \quad (20)$$

$\Delta\hat{\boldsymbol{y}}_t$ は，Nelder-Mead simplex 法 [9] などの反復法により求められる．

3.5　問題点

正弦関数摂動 von Mises 分布のモードは解析的に求まらない．そのため，式 (7) のように，位相・群遅延の両方を

ⓒ 2019 Information Processing Society of Japan

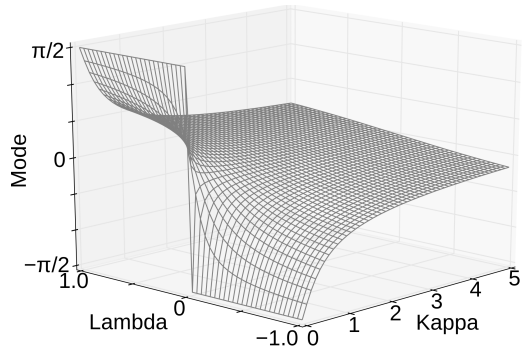

図 4　正弦関数摂動 von Mises 分布のモード

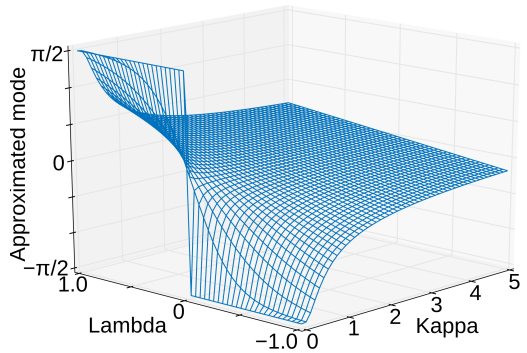

図 5　式 (22) により近似されたモード

考慮した目的関数の設計が困難[*1]である．

4. 正弦関数摂動 von Mises 分布のモード近似に基づく位相・群遅延モデリング

4.1 最尤推定解の分析と近似

ここで，正弦関数摂動 von Mises 分布の変数のモード

$$\hat{y} = \underset{y}{\operatorname{argmin}}\, P^{(\mathrm{ssvm})}\left(y; \mu, \kappa, \lambda\right) \tag{21}$$

を分析する．Nelder-Mead simplex 法を用いて，κ と λ に対する $\hat{y}$ を推定した結果を図 4 に示す．ただし，$\mu = 0$ としている．この図と式 (14) から，次のことが示される．

(1) μ は $\hat{y}$ に対してバイアスとして働く
(2) $\kappa = 0$ において，式 (1) は一様分布となり $\hat{y}$ は λ のみに依存
(3) κ が大きくなるにつれ，式 (1) は peaky な分布となり $\hat{y}$ は κ に強く依存
(4) $\lambda = 0$ において当該分布は式 (1) と等価となるため，$\hat{y} = \mu$

これらを踏まえ本稿では，$\hat{y}$ を次式で近似する．

$$\hat{y} \simeq f_{\mathrm{mle}}\left(\mu, \kappa, \lambda\right) = \frac{\pi}{2} \tanh\left(\frac{a\lambda}{\kappa^b + \epsilon}\right) + \mu \tag{22}$$

ここで a と b は係数である．これらの値は，Nelder-Mead simplex 法を用いて推定した $\hat{y}$ と近似値の二乗誤差が最小となるよう，数値計算により事前に推定する．ϵ は，値の発散を防ぐための微小値である．図 5 に，式 (22) で近似した $\hat{y}$ を示す．Nelder-Mead simplex 法を用いて反復的に推定した値（図 4）を良く近似していることを確認できる．

4.2 最大化する目的関数

式 (22) による近似を用いて，位相・群遅延のモデリングを行う．ここで，必要な情報を整理する．

(1) **位相モデリングには von Mises 分布 DNN が適切：** 位相の分布は一様分布に比較的近く，非対称性を有さないため，正弦関数摂動の導入は不要である．予備実験においても，摂動項の導入によるモデリング精度の向上は得られなかった．
(2) **式 (22) による近似を用いても，式 (7) と同様の定式化は困難：** 式 (7) は，位相の平均ベクトル $\boldsymbol{\mu}$ を $\boldsymbol{W}$ により線形変換することで，群遅延の分布のモード $\Delta\boldsymbol{\mu}$（平均と等価）を得る．この枠組みを，3 節の群遅延モデリングと式 (22) に適用すると，群遅延の分布のモードが与えられたもとでその分布の μ，κ，λ を推定する必要がある．これは不良設定問題であるため，その解に任意性が生じてしまう．

[*1] Deep NMF [10] のように，反復更新を neural network として展開することも手段の 1 つであるが，neural network 構造を複雑化してしまうため，本稿では採用しない．

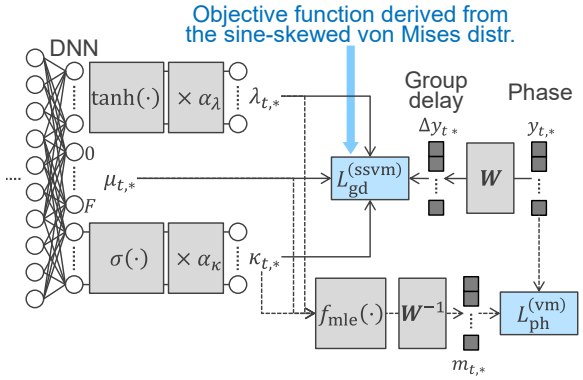

図 6　提案法における DNN の出力層から損失関数計算までのアーキテクチャ．$\sigma(\cdot)$ はシグモイド関数である．$\tanh(\cdot)$，$\sigma(\cdot)$，及び定数 α_λ と α_κ は，λ と κ の値を分布の定義域に収めるために利用される [7]．図中の $L_{\mathrm{ph}}^{(\mathrm{vm})}$ は，式 (28) の右辺の第一項に対応する．

以上を踏まえ，以下の目的関数を定義する．

$$L_{\mathrm{ph+gd}}^{(\mathrm{ssvm})} = -\log P^{(\mathrm{vm})}\left(\boldsymbol{y}_t; \boldsymbol{m}_t, \kappa\boldsymbol{I}\right) \cdot P^{(\mathrm{ssvm})}\left(\Delta\boldsymbol{y}_t; \boldsymbol{\mu}_t, \boldsymbol{\kappa}_t, \boldsymbol{\lambda}_t\right)^{\alpha_{\mathrm{gd}}} \tag{23}$$

ここで，

$$\boldsymbol{m}_t = \boldsymbol{W}^{-1}\Delta\boldsymbol{m}_t \tag{24}$$

$$\Delta\boldsymbol{m}_t = \left[\Delta m_{t,0}, \cdots, \Delta m_{t,f}, \cdots, \Delta m_{t,F}\right]^\top \tag{25}$$

$$\Delta m_{t,f} = f_{\mathrm{mle}}\left(\mu_{t,f}, \kappa_{t,f}, \lambda_{t,f}\right) \tag{26}$$

である．以降では，この目的関数に基づく学習・推定法を述べる．

4.3 学習

モデルパラメータ $\boldsymbol{\theta}_{\mathrm{G}}$ をもつ DNN は，図 6 に示すように，時刻 t において入力 $\boldsymbol{x}$ から推定値 $\boldsymbol{\mu}_t, \boldsymbol{\kappa}_t, \boldsymbol{\lambda}_t$ を出力する．次に，式 (24) を用いて，位相の von Mises 分布の平均 $\boldsymbol{m}_t$ を推定する．$\boldsymbol{\theta}_{\mathrm{G}}$ は，次式に示すように，式 (23) の負の対数値を最小化するように推定される．

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{G}} = \underset{\boldsymbol{\theta}_G}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} L_{\mathrm{ph+gd}}^{(\mathrm{ssvm})} \tag{27}$$

$$L_{\mathrm{ph+gd}}^{(\mathrm{ssvm})} = -\sum_{f=0}^{F} \cos\left(y_{t,f} - m_{t,f}\right) + L_{\mathrm{gd}}^{(\mathrm{ssvm})} \tag{28}$$

4.4 推定

学習済み DNN を用いて位相を推定する．時刻 t における推定位相 $\hat{\boldsymbol{y}}_t$ は次式で得られる．

$$\hat{\boldsymbol{y}}_t = \underset{\boldsymbol{y}_t}{\operatorname{argmin}} L_{\mathrm{ph+gd}}^{(\mathrm{ssvm})} \simeq \boldsymbol{m}_t \tag{29}$$

すなわち，学習済み DNN の出力から計算される群遅延の近似最尤推定解に $\boldsymbol{W}^{-1}$ をかけた値（位相の von Mises 分布の平均・モードと等価）が，推定位相となる．以上より，統一された目的関数 $L_{\mathrm{ph+gd}}^{(\mathrm{ssvm})}$ により，DNN 学習と位相推定が可能である．

5. 実験的評価

5.1 実験条件

実験的評価は，単一話者による読み上げ音声コーパス JSUT [11] を用いて実施した．学習データは，サブセット BASIC5000 に含まれる 5,000 文（約 6 時間），評価データは，サブセット ONOMATOPEE300 に含まれる 300 文である．サンプリング周波数は 16 kHz である．フレーム分析における窓長，シフト長，フーリエ変換長はそれぞれ，400 サンプル (25 ms)，80 サンプル (5 ms)，及び 512 サンプルとする．使用した窓関数は，ハミング窓である．DNN への入力特徴量は，当該フレーム及びその前後 2 フレームの対数振幅スペクトルを連結したベクトルである．入力特徴量は，学習時に平均 0・分散 1 に正規化する．DNN のアーキテクチャは，Feed-Forward 型であり，3 層・1024 ユニットの gated linear hidden unit [12] を持つ．DNN のモデルパラメータは乱数により初期化する．最適化アルゴリズムには，AdaGrad [13] を利用する．従来法と提案法の α_{gd} はそれぞれ，0.1 [5]，0.0001 とする．式 (22) における a，b，ϵ はそれぞれ，0.442374，0.814196，10^{-16} とする．

本稿では，以下の 2 手法の性能を比較する．

- **von Mises**: 式 (7) 及び 図 1 （従来法 [5]）
- **sine-skewed**: 式 (23) 及び 図 6（提案法）

従来法，提案法ともに，低周波数帯域における位相を DNN で推定し，残りの周波数の位相を乱数で与える．位相推定の周波数帯域は，従来法の性能が最も高かった 0–4 kHz（128 次元） [5] とする．また，DNN により位相を推定した後，Griffin-Lim 法 [14] により位相を補正する．補正のための反復回数は 100 とする．

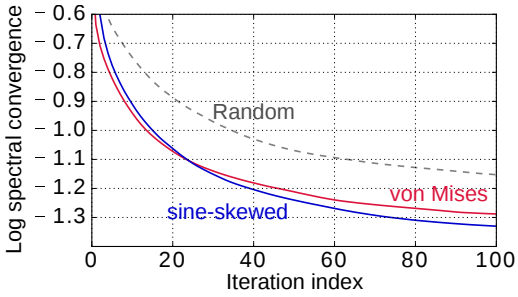

図 7 位相補正における spectral convergence の対数値の変化．この値が $-\infty$ となるとき，STFT と逆 STFT を通した完全再構成が成り立つ．この図は，評価データの一つの結果のみを示しているが，評価データの全てで同様の傾向が得られる．

表 1 音質に関する主観評価結果

Method A	Scores	p-value	Method B
von Mises	0.476 vs. 0.524	0.284	sine-skewed

5.2 客観評価

位相補正において従来法と提案法を比較する．図 7 に，位相補正の各反復回数における spectral convergence を示す．また，比較のため，ランダム値で初期位相を与えた結果（“Random”）を示す．[5] においても示されている通り，ランダム初期位相よりも，従来法は小さい spectral convergence を持つ．提案法の spectral convergence は，反復初期においては従来法よりも大きいものの，最終的に従来法より小さい値に収束していることが分かる．以上より，提案法は STFT ／逆 STFT による完全再構成に近い位相を生成できることが明らかである．

5.3 主観評価

提案法の有効性を確認するため，従来法と提案法による音声品質を比較する．比較のために，我々のクラウドソーシング型評価システムにおけるプリファレンス AB テストを実施した．25 人が参加し，各評価者に対し 50 円を支払った．評価者には，高音質の音声サンプルを選択させた．各手法の音声サンプルはランダムに提示した．表 1 に評価結果を示す．有意ではないが，提案法による音質の改善を確認できる．以上より，提案法の有効性が示される．

6. おわりに

本稿では，振幅スペクトログラムからの位相復元を目指し，群遅延モデリングに正弦関数摂動 von Mises 分布 DNN を用いた DNN 学習・位相推定法を提案した．正弦関数摂動 von Mises 分布のモードを微分可能な形で近似することで，学習・推定で矛盾のない目的関数を設計し，実験的評

ⓒ 2019 Information Processing Society of Japan

価から，従来法よりも小さい spectral convergence をもつことを明らかにした．今後は，音声以外における提案法の有効性を確認する．

謝辞： 本研究の一部は，セコム科学技術支援財団，総務省 SCOPE の助成を受け実施した．

参考文献

[1] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] S. Takaki, H. Kameoka, and J. Yamagishi, “Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis,” in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” vol. abs/1609.03499, 2017. [Online]. Available: https://arxiv.org/abs/1703.10135

[4] Y. Saito, S. Takamichi, and H. Saruwatari, “Text-to-speech synthesis using stft spectra based on low-/multi-resolution generative adversarial networks,” in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5299–5303.

[5] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, “Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network,” in *Proc. IWAENC*, Tokyo, Japan, Sep. 2018.

[6] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons Ltd., 1999.

[7] 高道慎之介, 齋藤佑樹, 高宗典玄, 北村大地, and 猿渡洋, “方向統計 DNN に基づく振幅スペクトログラムからの位相復元,” in **日本音響学会** *2018* **年秋季研究発表会**, 大分, Sep. 2018.

[8] T. Abe and A. Pewsey, “Sine-skewed circular distributions,” *Statistical Papers*, vol. 52, no. 3, pp. 683–707, Aug. 2011.

[9] J. A. Nelder and R. Mead, “A simplex method for function minimization,” *The Computer Journal*, pp. 308–131, Jan. 1965.

[10] J. L. Roux, J. R. Hershey, and F. Weninger, “Deep NMF for speech separation,” in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 66–70.

[11] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis,” vol. abs/1711.00354, 2017.

[12] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” vol. abs/1612.08083, 2016.

[13] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *EURASIP Journal on Applied Signal Processing*, vol. 12, pp. 2121–2159, 2011.

[14] D. W. Griffin and J. S. Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

ⓒ 2019 Information Processing Society of Japan